Practical Considerations in Choosing between the Case-Cohort and Nested Case-Control Designs

Sholom Wacholder

When the outcome is rare, both the nested case-control and case-cohort designs can provide economical estimates of relative risk parameters under the proportional hazards model, while requiring exposure and covariate information on only a small subset of the cohort. Comparisons of the statistical efficiency of the case-cohort and nested case-control designs for studies with substantial late entry or censoring suggest a small to moderate advantage for the case-control study[1-3]; in contrast, the results for studies with little late entry or censoring[1,4,5] slightly favor the case-cohort design. Below, I discuss some practical points that have been relevant in several studies of etiologic factors for cancer where the analysis requires tight control or matching on a time variable.

When control for time is not essential, as in studies of perinatal mortality, the case-cohort design is identical to the *case-base* design.[6,7] The primary consideration for choosing between a case-control and a case-base design seems to be whether the odds ratio or the risk ratio is regarded as the primary parameter of interest.[8,9]

Case-Cohort or Nested Case-Control Design?

The case-cohort design is an unmatched variant of the nested case-control design.[10,11] In the case-cohort design, a *subcohort,* which is a simple random sample of the cohort, is the source of all controls, while the case-control design selects controls that are matched to the cases on time. Apart from details of analysis, the case-cohort design is inherently simpler than the case-control design. The advantages of the case-cohort design among the points mentioned below are consequences of the random sampling, while those of the nested case-control design are due to matching on time.

Ease of Analysis

The case-control design and analysis are widely known by epidemiologists. Computer software for analysis is widely available. Difficulty of analysis, however, will no longer be a major impediment to the use of the case-cohort study design as software for analysis of case-cohort studies becomes available. For example, the IBM-PC package EPITOME, a forthcoming National Cancer Institute publication by John Boice, Jay Lubin, and Dale Preston, contains software for analysis of case-cohort data.

Ease of Planning

The power of the case-cohort design depends on the size of the overlap in the sets of subjects at risk at each pair of event times.[3] Therefore, the power cannot be calculated easily unless the cohort is *assembled,*[3] that is, the cohort members are identified, their dates of entry to and exit from the cohort are known, and the dates of diagnosis for all the cases are known. On the other hand, the power of the nested case-control study is virtually independent of the size of the cohort.[3,12]

Multiple Outcomes

A key advantage of the case-cohort design is the ability to use the same subcohort for several diseases[1,2,4,6,11] or for subtypes of disease, such as different forms of leukemia. Langholz and Thomas suggest that this may require adjustment of significance levels and confidence intervals "to account for the induced correlation between outcomes."[2, page 174] But no adjustment needs to be made in the analysis when the focus of the investigation is on the evaluation of risk factors for each disease separately, rather than on the comparison of risk factors for different diseases.

External Comparisons

In the case-cohort design, the subcohort is a simple random sample of the entire cohort. This enables simple estimation and modeling of the absolute covariate-specific incidence[4] and the use of standardized mortality ratios[11] to allow comparisons of disease incidence in the cohort to that of the general population.[13] External comparisons can be useful when the general population is almost completely unexposed, while nearly everyone in the cohort is exposed. Lubin and coauthors, in an unpublished manuscript, generalize the approach Boivin and I[11] proposed by showing how Poisson regression models[12] can be used to make external comparisons from case-cohort data. These methods can be applied, for example, in studies of second cancers after treatment for

From the National Cancer Institute, Biostatistics Branch, 6130 Executive Blvd., EPN 403, Bethesda, MD 20892.

non-Hodgkin's lymphoma, where nearly everyone in the cohort has been treated with either radiation or chemotherapy, the exposures of primary interest, but only a small fraction of the general population has been exposed. The standardized mortality ratio can be estimated from nested case-control data[14]; however, there can be bias when the ratio of controls to cases is small.[14]

Multiple Time Scales

The nested case-control design includes matching on a particular time variable, such as age or time since entry into the cohort. Therefore, all the analyses must use that variable as the primary time scale. In contrast, in the case-cohort design, the investigators can choose the appropriate time scale for each analysis. For example, in a study of second primary cancers after an initial diagnosis of oral cancer, a time scale of *age*[15] would be appropriate if *time since first cancer* were unrelated or weakly related to the risk of a solid tumor at another site. But a time scale of *time since first cancer* might be appropriate for quantifying the effect of quitting smoking during the year after diagnosis of the first cancer. On the other hand, that time scale may result in overmatching for estimating the effect of *time since quitting smoking* on disease risk. To take an extreme case, if all smokers quit immediately upon diagnosis of the first primary, there would be no variability in *time since quitting* in the matched sets and therefore zero power to detect an effect of *time since quitting.* But if *age* were the time scale in this example, *time since quitting* would be completely confounded by *time since first cancer;* this would also be troubling unless an effect of *time since first cancer* can be ruled out. A nested case-control design would include matching on either *age* or *time since first cancer,* restricting the options of the investigators at the analysis stage. In a case-cohort study, all analyses available in a full cohort study can be performed; some could use *time since first cancer,* while others could use *age* as the primary time scale.

Time-Dependent Exposure

An advantage of the nested case-control study is that information on time-dependent exposures in cases and controls does not need to be collected beyond the time of follow-up of the case. This feature can generate a major savings in the time required for abstraction of medical records for detailed chemotherapy history, for example. When the marginal effort required to extend the time period for which exposure history is gathered is small, however, it may be worthwhile to collect it without a time cutoff to reduce the possibility of the costly error of not gathering some of the exposure history that is needed for the analysis.

Future Follow-up

The possibility of further follow-up on the same cohort has implications on the choice of design. In a case-cohort study, the same subcohort can be used for a period of extended follow-up. New cases found during the extended follow-up do not require identification of new controls, only ascertainment of additional exposure since the end of the first follow-up in subcohort members who were previously identified and are at risk beyond the end of the initial follow-up. In a nested case-control study, there is no need to update exposure for controls selected previously; on the other hand, each new case requires the effort of ascertaining complete exposure history for new controls. Because the time elapsed since the beginning of exposure has increased, substantial effort may be required to obtain accurate and complete exposure information for recently selected controls. Because fewer subjects need to be studied, this consideration favors the case-control design unless there is extra difficulty in ascertaining exposure from the earlier period. If, for example, recent chemotherapy records are computerized while older ones require abstraction from paper records, the case-cohort design has an advantage because the older records will not be needed when the exposures of subcohort members are updated. Neither design obviates the effort to follow the entire cohort for disease experience or to ascertain complete exposure histories on the newly identified cases. Care must be taken to avoid differential misclassification if cases' records are found and abstracted at a later time than controls'.[2]

Secondary Use of Controls

The subcohort from a case-cohort study can also be used for other purposes. For example, Prentice[4] suggested that subcohort members could be used for monitoring compliance to a treatment in a clinical trial.

We plan to exploit this advantage in a case-cohort study of the relation between human papillomavirus infection and cervical neoplasia. Each woman in a cohort will be screened annually for neoplasia via a Papanicolaou smear and cervicography; at the same time, exfoliated cervical cells will be collected to assay for presence of human papillomavirus DNA. Unfortunately, we may not be able to afford to assay all the specimens for all the women; however, since the assay works equally well with frozen cells, the biological materials can be stored and made available for assay at a later time. Although both designs could achieve the study's primary goal, the availability of a subcohort makes the case-cohort design preferable to a nested case-control design:

The subcohort can be used to estimate the prevalence of viral infection in the cohort, and the general population, from which the cohort itself was selected as a random sample.
Specimens from subjects in the subcohort can be assayed immediately after they are obtained for use in longitudinal studies of changes in viral status.
We plan to collect exfoliated cells from a subset of the subcohort at 3-month intervals to learn more about short-term persistence of viral infection.
The subcohort can be used as a source of controls for a case-control study of prevalent neoplasia.

Rapidity of Control Selection

Control selection for a nested case-control study must await identification and confirmation of the cases; however, in the case-cohort design, selection of controls is independent of characteristics of the cases and therefore can begin immediately.[4,11] This difference can result in faster completion of data collection for a case-cohort design. For example, in a multi-city study of an occupational exposure, a cohort of around 100,000 subjects, including perhaps 100 cases who work in factories with the exposure, is being identified and followed for new cases of cancer. We were planning a nested case-control study that would obtain detailed exposure measurements. We were interested in completing the fieldwork for the study as quickly as possible. Identification of cases and the other members of the cohort proceeded slowly, however, and we considered switching to a case-cohort study because selection of the subcohort could begin immediately. As members of the entire cohort were identified, subcohort members could be sampled with a fixed fraction p_1, leading to a more rapid completion of the study.[11] In the case-control design, if controls were selected for a given case before complete identification of the roster of cohort members, it would be necessary to sample from the additional subjects so that all eligible controls have an equal chance to be selected. While this is feasible, it can be a complex and time-consuming task, resulting in a total number of controls that is larger than anticipated.

If more cases than anticipated are identified or if the cohort is smaller than expected, it is easy to adjust the total subcohort size during the course of a case-cohort study. Begin with a sampling fraction p_1 that is smaller than what is anticipated to be the final sampling fraction. As the study proceeds and more refined estimates of the numbers of cases and cohort members are obtained, the final desired sampling fraction p_2 can be set. If $p_2 > p_1$, newly identified cohort members can be selected for the subcohort with a sampling fraction of p_2. Previously identified cohort members who were given a chance but were not selected will need an additional chance equal to $(p_2 - p_1)/(1 - p_1)$ to be selected into the subcohort so that everyone has the same overall chance p_2 of being sampled.

The independence between control selection and attributes of the cases in the case-cohort design can also be helpful at the end of the study. When a case is identified during the closing days of the study, exposure ascertainment is needed only for that case. For a case-control study, additional effort is required for identification of several controls and ascertainment of their exposure. Similarly, in the case-control design, one must either wait for a pathology report confirming the diagnosis of a case or risk wasting the effort of exposure ascertainment for the controls.

Discussion

In a particular situation, some of the points discussed above may outweigh statistical efficiency, particularly in the absence of a major difference, in the choice between designs. The importance that should be given to each point, and to statistical efficiency, should depend on factors that are specific to the study, including the objectives of the research, how the cohort and the cases within the cohort will be identified, and how covariate information can be obtained.

References

1. Wacholder S, Gail MH, Pee D, Brookmeyer R. Alternative variance and efficiency calculations for the case cohort design. Biometrika 1989;76:117–123.
2. Langholz B, Thomas DC. Nested case-control and case-cohort methods of sampling from a cohort: a critical comparison. Am J Epidemiol 1990;131:169–176.
3. Wacholder S, Gail MH, Pee D. Selecting an efficient design for assessing exposure-disease relationships in an assembled cohort. Biometrics (in press.)
4. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986;73:1–11.
5. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case cohort studies. Ann Stat 1988;16:64–81.
6. Kupper LL, McMichael AJ, Spirtas R. A hybrid epidemiologic study design useful in estimating relative risk. J Am Stat Assoc 1975;70: 524–528.
7. Miettinen OS. Design options in epidemiologic research. an update. Scand J Work Environ Health 1982;8(Suppl. 1):7–14.

8. Greenland S. Adjustment of rate-ratios in case-base studies (hybrid epidemiologic designs). Stat Med 1986;5:579–584.
9. Flanders WD, Dersimonian R, Rhodes P. Estimation of risk ratios in case-base studies with competing risk. Stat Med 1990;9:423–435.
10. Mantel N. Synthetic retrospective studies and related topics. Biometrics 1973;29:479–486.
11. Wacholder S, Boivin J-F. External comparisons with the case-cohort design. Am J Epidemiol 1987;126:1198–1209.
12. Breslow NE, Day NE. Statistical Methods in Cancer Research, Vol. 2, The Design and Analysis of Cohort Studies (IARC Scientific Publications No. 82). Oxford: Oxford University Press, 1987.
13. Bergkvist L, Adami H-O, Persson I, Hoover R, Schairer C. The risk of breast cancer after estrogen and estrogen-progestin replacement. N Engl J Med 1989;321:293–297.
14. Breslow N, Langholz B. Nonparametric estimation of comparative mortality functions. J Chron Dis 1978;40(Suppl. 2):89S–100S.
15. Breslow NE, Lubin JH, Marek P, Langholz B. Multiplicative models and cohort analysis. J Am Stat Assoc 1983;78:1–12.